基于判别特征加权的GPLVM算法

王秀美 高新波
(西安电子科技大学电子工程学院 西安 710071)

摘 要 高斯过程隐变量模型是最近提出的比较流行的无监督降维方法。但是,它是一种无监督的机器学习算法,没有突出类结构,使得结果不能有效地表示类别信息。因此,提出一种利用判别特征值对高斯过程隐变量模型进行加权的算法,该算法不仅能够加强模型在低维流形上的判别性,而且能很好地保持类内的流形结构。

关键词 高斯过程隐变量模型,因子分析,概率主成分分析,局部Fisher判别分析

Weighted GPLVM Algorithm Based on Discriminant Features

WANG Xiu-mei GAO Xin-bo
(School of Electronic Engineering, Xidian University, Xi'an 710071, China)

Abstract Gaussian process latent variable model (GPLVM) is a popular manifold method recently proposed for dimensional reduction. However it cannot keep some class structure of datasets for it is an unsupervised learning method. A weighted GPLVM algorithm will be given using the discriminant features. This algorithm can approve an discrimiant results and keep good manifold in each class.

Keywords Gaussian process latent variable model, Factor analysis, Probabilistic principal component analysis, Local Fisher discriminant analysis

1. 引言

"维数灾难"是统计和机器学习的常见问题,在人脸识别、语音识别、动态跟踪等实际问题中,数据维数过高,使得在数据处理中计算量很大,不能对数据进行有效的处理。因此,寻求较好的降维方法,以正确获得高维数据集的潜在低维结构信息,是当前大家普遍关注的问题。

目前提出的降维方法有很多,其中线性降维方法因为其思想简单,易于实现而得到了广泛应用。因子分析[1](Factor Analysis, FA)和主成分分析[2,3](PCA)、线性判别分析[4](LDA)等线性降维方法在各个领域都发挥着重要的作用,这些方法利用的是样本在特征空间的不同特征。PCA的主要思想是用较少的综合变量来代替原来较多的变量,同时要求这几个综合变量互相不相关,并能尽可能多地表示原来数据的能量,它是将多指标化为少数几个综合指标的一种统计分析方法。由PCA衍生并推广出的概率主成分分析(Probabilistic PCA, PPCA)和核主成分分析(Kernel PCA, KPCA)两种降维方法,前者PPCA是对PCA的概率推广,仍然是通过线性变换达到降维的目的;后者KPCA则是对降维进行了非线性扩展。

但是,上述降维方法都有自身的不足。LDA侧重在样本的分类性,但不能保持数据的能量信息。PCA没有突出类间的结构。且PCA和PPCA都是线性降维方法,虽然易于实施,但是对一些结构复杂的数据却无能为力,这是由线性方法本身的限制性造成的。KPCA虽然将PCA进行了非线性推广,但是它只给出了如何从观测空间到低维隐空间的映射。如何从低维到高维映射建立联系,KPCA没有给出解决的办法[5]。更为重要的是,以上方法只考虑了样本能量特征,因此需要大量的样本数据,对于小样本数据,往往不能得到满意的结果。

针对上述方法的不足,研究人员提出了高斯过程隐变量模型(GPLVM),它是一种非线性降维技术[5,6],是由PPCA推导演变得到的。GPLVM建立了从隐变量空间到观察空间的非线性高斯过程映射,估计数据点的联合密度和它们在隐变量空间中的坐标位置,然后得到低维隐空间中的表达。它克服了线性降维方法的局限,并且建立了从低维隐空间到高维空间的映射关系。GPLVM有一个显著的特性,即当观察数据的样本比较少时,仍然可以用来寻找观察数据的低维流形,也就是说GPLVM非常适合处理小样本的高维数据[7,8]。

然而,GPLVM所用到的高斯过程映射是在样本的每一维上单独进行,是无监督的,不能很明显地突出一些类结构信息。因此,本文提出了改进的GPLVM,也可以称为加权的高斯隐变量模型(W-GPLVM),利用实际数据中已知数据的监督信息(如样本类标)通过变换,得到每一维上的判别特征值,然后用这些特征去加权相对应维数上的联合密度分布。它对寻求低维空间上的流形有很显著的帮助,而且能有效地增强

到稿日期:2008-06-10 本文受新世纪优秀人才支持计划(No. NCET-04-0948),教育部长江学者和创新团队支持计划(No. IRT0645),国家自然科学基金(No. 60702061)资助。

王秀美 博士生,研究方向是统计机器学习、概率模型、贝叶斯学习,E-mail:wangxiumei@gmail.com;**高新波** 教授,博士生导师,IEEE高级会员,主要研究方向是影像处理、分析和理解、模式识别和机器学习。

GPLVM 的判别性。

2 高斯过程隐变量模型(GPLVM)

假设 $Y=[y_1,\cdots,y_N]^T$ 代表观测的数据矩阵,其中 $y_i\in R^D$,N 为样本个数。$X=[x_1,\cdots,x_N]^T$ 表示隐变量空间中的数据矩阵,其中 $x_i\in R^d$。高斯隐变量模型是建立由隐变量空间 X 到观测空间 Y 的非线性映射,以 X 为参数矩阵,求保证观测数据 Y 的联合密度最大时,参数 X 的值,即确定观测数据对应在隐空间中的坐标。高斯隐变量模型是对 PPCA 的推广。

2.1 概率主成分分析(PPCA)

PPCA 是一种概率降维方法[2],它建立了一种从低维空间(隐变量空间)到高维空间的线性映射。当考虑噪声时,它就不能保持线性关系,这里假设噪声 $\eta_n\in R^D$ 符合独立高斯分布。从低维隐空间到观测空间的映射可以表示为:

$$y_n=Wx_n+\eta_n \tag{1}$$

线性映射由 $W\in R^{D\times d}$ 确定,式(1)中的噪声分布如下式:

$$p(\eta_n)=N(\eta_n|0,\beta^{-1}I) \tag{2}$$

根据式(2)可以得到观测数据点的似然概率:

$$p(y_n|x_n,W,\beta)=N(y_n|Wx_n,\beta^{-1}I) \tag{3}$$

这时,假设隐变量空间中的点的先验概率分布为:

$$p(x_n)=N(x_n|0,I) \tag{4}$$

即假设隐变量空间中的点是独立同分布的。通过对隐变量空间中的点积分可以推导出边缘似然:

$$\begin{aligned}P(y_n\mid W,\beta)&=\int p(y_n\mid x_n,W,\beta^{-1}I)p(x_n)\,\mathrm{d}x_n\\&=N(y_n\mid WW^T+\beta^{-1}I)\end{aligned} \tag{5}$$

和观测数据的联合似然:

$$P(Y|W,\beta)=\prod_{n=1}^{N}p(y_n|W,\beta) \tag{6}$$

用共轭尺度梯度算法对式(6)求最大,就可以得到映射矩阵 W,进而可以得到观测点在隐空间中对应点的坐标。详细的证明步骤可以参考文献[2]。

2.2 高斯过程隐变量模型(GPLVM)

如上所述,GPLVM 是 PPCA 的推广。基于 PPCA 的推导过程,本节给出 GPLVM 的推导过程。在上面的推导中给出了一种假设:隐空间中的点是独立标准高斯分布,即式(4)。将式(4)代入式(3),对 x_n 求积分,可以得到边缘似然,从而可以推导出联合似然,如式(6)所示。

GPLVM 算法是将隐空间中点的坐标 X 作为参数,通过使观察数据联合密度用共轭尺度梯度法最大,来确定低维空间中点的坐标。这时就对式(3)中的另一个未知量——投影矩阵 W 可以做如下假设:

$$p(W)=\prod_{i=1}^{D}N(w_i|0.I) \tag{7}$$

式(7)表示投影矩阵 W 在每一维上的投影 w_i,$i=1,\cdots,D$ 是独立同分布的。对 W 求得边缘似然如下式:

$$P(y_{:,d}|X,\beta)=N(y_{:,d}|XX^T+\beta^{-1}) \tag{8}$$

则联合似然为:

$$P(Y|X,\beta)=\prod_{i=1}^{D}\frac{1}{(2\pi)^{\frac{N}{2}}|K|^{\frac{1}{2}}}\exp(-\frac{1}{2}y_{:,i}^TK^{-1}y_{:,i}) \tag{9}$$

这里 $K=XX^T+\beta^{-1}I$。由此可见,此时从低维空间到高维空间的映射是线性的。若假设该映射为 f,并使 f 满足一个高斯过程的先验:

$$p(f|K)=N(f|0,K) \tag{10}$$

此时 f 是非线性的,因此从低维到高维的映射就建立起非线性映射。

对 f 积分,得边缘似然:

$$\begin{aligned}p(y_{:,d}\mid X,\beta)&=\int p(y_{:,d}\mid x_n,f,\beta^{-1}I)p(f)\mathrm{d}f\\&=N(y_{:,d}\mid 0,K)\end{aligned} \tag{11}$$

和观测数据的联合似然分布:

$$P(Y\mid X,\beta)=\prod_{d=1}^{D}(y_{:,d}\mid X,\beta)=\prod_{d=1}^{D}\frac{1}{(2\pi)^{\frac{N}{2}}|K|^{\frac{1}{2}}}\exp(-\frac{1}{2}y_{:,d}^TK^{-1}y_{:,d}) \tag{12}$$

其中,K 是协方差函数矩阵,或者称为核函数矩阵。本文取得核函数为:

$$k(x_i,x_j)=\theta_{rbf}\exp(-\frac{\gamma}{2}(x_i-x_j)^T(x_i-x_j))+\theta_{bias}+\theta_{white}\delta_{ij} \tag{13}$$

$k(x_i,x_j)$ 为矩阵 K 的第 i 行第 j 列对应的元素。δ_{ij} 为 Kronecker delta 函数。这时从隐空间到高维空间的映射是一个非线性映射的高斯过程,而且对于观测数据,每一维是独立的。联合似然式(9)可以进一步简化为:

$$P(Y\mid X,\beta)=\frac{1}{(2\pi)^{\frac{N}{2}}|K|^{\frac{1}{2}}}\exp(-\frac{1}{2}tr(K^{-1}YY^T)) \tag{14}$$

3 加权 GPLVM 算法

由式(9)可以看出,观察数据的联合密度分布是每一维上似然的乘积。而且,各维度的表达式完全相同,每个维度对于求联合似然的贡献是相同的。实际问题中,观察到的数据在一定程度上反映的信息会有所重叠,在高维空间中研究样本的规律也比较复杂,因此必须根据实际需要对高维数据进行降维,GPLVM 无疑是一种较好的方法。

在用 GPLVM 算法时,如果已经知道训练数据的一些类标信息,那么可以利用一些判别分析的方法,找到对判别保持比较好的维度,然后对这些维度进行加权,以增强算法的区分性。本文利用 LFDA 提取判别特征,然后对 GPLVM 进行加权。

预先给定观察数据 Y,类标 $l_i\in(1,\cdots,L)$,其中 $|l_i|=n_i$,$\sum_{i=1}^{L}n_i=n$,由此可以得到类内散度矩阵 S_w,类间散度矩阵 S_b,

$$\begin{aligned}S_b&=\frac{\sum_{k=1}^{N}n_k(m_k-m)(m_k-m)^T}{N}\\S_w&=\frac{\sum_{k=1}^{N}\sum_{i\in l_k}(y_i-m_k)(y_i-m_k)^T}{N}\end{aligned} \tag{15}$$

其中 m_k 表示类标为 k 的样本的均值,m 是指所有样本的均值:

$$m_k=\frac{1}{n_k}\sum_{i\in l_k}y_i,m=\frac{1}{N}\sum_{i=1}^{N}y_i \tag{16}$$

对于常见的 Fisher 判别分析来说,就是要找到投影矩阵 $T=(\tau_1,\cdots,\tau_D),\tau_i\in R^D,(i=1,\cdots,D)$,使得

$$T=\underset{T\in R^{D\times D}}{\operatorname{argmax}}tr((T^TS_wT)^{-1}(T^TS_bT)) \tag{17}$$

而 τ_i 对应于特征 λ_i(其中 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_D$)得到:

$$S_b\tau_i=\lambda_iS_w\tau_i,i=1,\cdots,D \tag{18}$$

λ_i 是投影向量 τ_i 对应的特征向量,它代表在第 i 维上判别信

息能量的强弱。

相对于传统的 Fisher 判别分析，局部 Fisher 判别分析有它特有的性质，它通过对类间散度矩阵和类内散度矩阵进行改进，更好地保持了数据的局部结构。详细的推导和描述参考文献[9]。

本文利用 LFDA 可以得到投影矩阵和特征值，对原始的观察数据先用投影矩阵 $T=(\tau_1,\cdots,\tau_D)$ 作用，得到 Y'，再用特征值对相应维数的似然进行加权，就可以得到 W-GPLVM 算法。只有先用投影矩阵对原始数据进行变换，才能保证特征值和投影向量作用的维度相互对应。需要注意的是，这里利用判别分析方法只是想得到各个维数上对应的判别能量的大小，而不是用来降维。

最终得到加权的似然为：

$$\tilde{P}(Y'\mid X,\beta)=\prod_{d=1}^{D}\lambda_d p(y'_{:,d}\mid X,\beta)=\prod_{d=1}^{D}\lambda_d\frac{1}{(2\pi)^{N/2}\mid K\mid^{1/2}}$$

$$\exp\left(-\frac{1}{2}y'^{T}_{:,d}K^{-1}y'_{:,d}\right)\qquad(19)$$

4 实验结果分析

在实验中，用两组数据比较 W-GPLVM 和 GPLVM 的性能。用局部 Fisher 判别分析(LFDA)的方法来提取加权的特征，同时将观察数据在相对应的投影矩阵上投影，以保证是对相对应的维度进行了特征加权。实验中，所得的低维空间为两维，以便能够对比较的结果有更直观、更形象的认识。

实验中采用了两类数据，一类是石油数据，同一管道中流出的石油由 12 个指标来表示，即数据有 12 维，分 3 类表示不同的油质，共有 1000 个样本，数据详细说明见文献[10]，实验结果如图 1 所示。另外一类是手写数字(hand-written digits)，其每个数字是 256 维，每次实验取单个数字图像的 600 个样本。对于第二类数据，我们取 3 组。每一组我们都选取难以区分的数组，例如第一组是数字 3 和 5；第二组是 3，6 和 9；第三组是 2，4，6。分别表示在图 2、图 3 和图 4 三组图中。

两类数据共有 4 组，从上到下分为 4 行。第一组是对第一类数据，后面 3 组是第二类数据。每一组中，从左至右排列了 3 幅图像。图(a)表示原始的 GPLVM 算法的结果；图(b)是对观察数据先用 LFDA 然后用 GPLVM 算法的结果；图(c)为本文所提出的算法 W-GPLVM 的结果。每幅图中不同的颜色表示不同类型的数据。

从以上 4 组实验结果可以看出，LFDA＋GPLVM 比 GPLVM 更能区分每一类的结果，但是它却丢失了 GPLVM 模型本身的一个重要特性，那就是找出高维数据的流形结构。图 1 所示的石油数据分为 3 类，与前两种算法结果比较，本文提出的算法既使得每一类数据比较集中，又保持了类内的流形趋势。第二类数据实验中，选取的手写数字图片的维数是 256，降至两维时，仍然维持了原来数据的流形性质，并且相对于 GPLVM 算法的实验结果，又有很好的分类性。

综上，GPLVM 判别性较差，LFDA＋GPLVM 虽能进行较好的分类，但是却不能对数据的类内结构有一个很好的表示。与前两种算法相比，W-GPLVM 算法更具优势：除了类间区分性比较明显外，类内的流形结构也容易看出，原因在于它一方面能利用先验的类标知识，在低维上对数据表达时保持了判别性，另外一方面它对高维数据的流形结构进行了一个比较明显的表达。

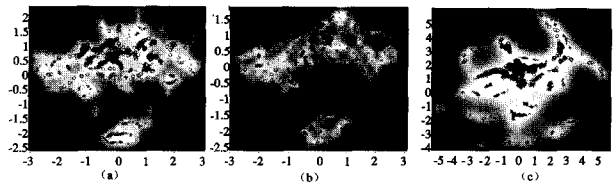

图 1

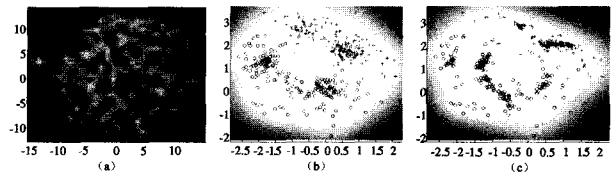

图 2

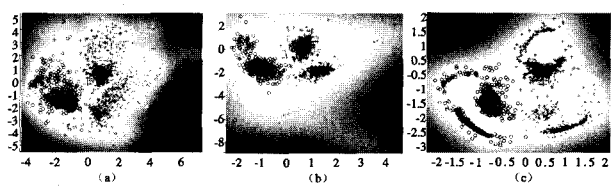

图 3

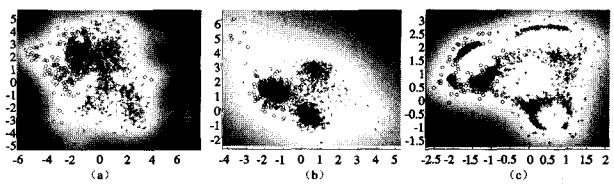

图 4

结束语 本文将局部 Fisher 判别分析方法和高斯过程隐变量模型相结合，利用局部 Fisher 判别分析方法得到判别向量及其所对应的特征值，并将特征值对高斯过程因变量模型加权，把监督信息用到无监督高斯过程因变量模型算法。相比于高斯过程因变量模型，加权后的算法既可以对高维数据进行有效的降维，得到的结果具有区分性类信息，在每一类内又有比较明显的流形结构。

本文中用到的监督信息为每一类的类标，除此之外，如果有其它的监督信息，参照生成图的方法，可以考虑建立有监督或者半监督的 GPLVM 算法[10,11]。

参 考 文 献

[1] Bartholomew D J. Statistical factor analysis and related methods. New York: Wiley, 2004

[2] Tipping M E, Bishop C M. Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B, 1999, 61(3): 611-622

[3] Roweis S T. EM algorithms for PCA and SPCA. Advances in Neural Information Proce-ssing Systems, The MIT Press, 1997, 10: 626-632

[4] Hastie T, Tibshirani R. Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society, series B, 1996, 58: 158-176

[5] Lawrence N D. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. Journal of

Machine Learning Research, 2005, 6: 1783-1816
[6] Lawrence N D. Gaussian process models for visualization for high dimensional data. Advances in Neural Information Processing Systems, The MIT Press, 2004, 16: 329-336
[7] Lawrence N D. Learning for large datasets with the Gaussian process latent variable models. Advances in Neural Information Processing Systems, The MIT Press, 2004, 16: 329-336
[8] Wang J M, Fleet D J, Hertzmann A. Gaussian process dynamical models for human motion. IEEE Trans. Pattern Recognition and machine Intelligence, 2008, 30(2): 283-298
[9] Sugiyama M. Local Fisher discriminant analysis for supervised dimensionality reduction // Proceedings in the 23rd International Conference on Machine Learning. Pittsburgh, PA, 2006
[10] Bishop C M, James G D. Analysis of multiphase flows using dual-energy gamma densitometry and neural networks. Nuclear Instruments and Methods in Physics Research, series A, 1993, 327: 580-593
[11] Urtasun R, Darrell T. Discriminant Gaussian process latent variable model for classification // Proceedings in the 24th International Conference on Machine Learning. Corvallis, OR. USA, 2007

(上接第 125 页)

其中，$M(i,j)$表示规划化矩阵中主体服务 AS_{pi} 的第 j 项 QoS 属性规格化值，$Q_{max}(j)$表示第 j 项 QoS 属性值中的最大值，$Q_{min}(j)$表示第 j 项 QoS 属性值中的最小值，$Q(i,j)$表示主体服务 AS_{pi} 第 j 项 QoS 属性值。

步骤 2　通过式(10)计算各个主体的总体 QoS 水平。

$$QoS(AS_{pi}) = \sum_{j=1}^{4} (\mu_j \times M(i,j) \times fid(j)) \quad (10)$$

其中，μ_j 表示第 j 项 QoS 属性对应的权重，$0 \leqslant \mu_j \leqslant 1$，$\sum_{j=1}^{4} \mu_j = 1$，$fid(j)$表示第 j 项 QoS 属性对应的真实度，$0 \leqslant fid(j) \leqslant 1$。

4.3 算法分析

定理 1　支持 QoS 约束的语义主体服务匹配算法是有效的，即该算法的返回结果能满足主体服务请求的要求。

证明：本文提出的算法是一种松弛匹配算法，由主体服务语义相似度匹配过程可知，在语义相似度阈值范围内，可以有效找到通用描述和功能描述相似语义的主体服务。服务质量匹配过程是在上述主体服务中获得总体水平最高的一个。因此，本文算法返回结果能满足语义和服务质量要求的主体服务请求。证毕。

定理 2　支持 QoS 约束的语义主体服务匹配算法的时间复杂度是 $O(n^3)$。

证明：采用 Dijstra 最短路径算法进行几何距离计算的时间复杂度是 $O(n^2)$，因此，主体服务语义相似度匹配的时间复杂度是 $O(n^3)$。规格化矩阵计算的复杂度是 $O(4 \times |SAS_p'|)$，其中 $|SAS_p'|$ 是满足相似度匹配的主体服务数量，通常为常数。因此，整个算法的时间复杂度为 $O(n^3)$。证毕。

结束语　本文研究了自主单元的主体服务匹配问题，给出了自主单元主体服务描述模型。该模型能全面描述主体服务的功能语义，客观反映主体服务的服务质量特征。在此基础上，本文提出一种支持 QoS 约束的自主单元语义服务匹配算法。该算法首先通过语义相似度匹配，综合计算主体服务通用描述和功能描述的语义相似度；并在所有满足给定相似度阈值的主体服务中，利用简单加权法思想计算每个主体的总体服务质量水平，最终求得满足主体服务请求描述的最佳主体服务。

本文提出的主体服务匹配算法具有以下特征：(1)综合考虑了语义和服务质量因素，既解决了传统语法级服务描述的异构性，提高了查准率，又能充分体现出服务请求者对主体服务性能的不同需求，选择出最佳服务；(2)在语义相似度匹配中，将概念信息容量引入基于几何距离的方法中，并考虑边的强度，从而综合了基于几何距离和基于信息容量两种方法的优点，有效解决了同等条件下的概念难以区分、匹配效果不佳的缺点；(3)提出的服务质量模型和匹配算法简单有效，不限制 QoS 属性的类型、数量以及值的大小，可扩展性和灵活性好；(4)在计算每个主体服务的总体服务质量水平过程中，在简单加权和规格化思想基础上，引入每个质量属性的真实度因素，全面、客观反映各个主体服务的服务质量性能。

进一步的研究工作是提高服务匹配算法的效率以及在模糊条件下的主体服务匹配。

参 考 文 献

[1] Kephart J, Chess D. The Vision of Autonomic Computing [J]. IEEE Computer Society, 2003, 1: 41-59
[2] Gerald T, David M C, William E, et al. A Multi-agent Systems Approach to Autonomic Computing [C] // Proceedings of AAMAS'04. New York, USA, July 2004
[3] Tom D W, Tom H. Towards Autonomic Computing: Agent-based Modeling, Dynamical Systems Analysis, and Decentralized Control [J] // Proceedings of IEEE International Conference on Industrial Informatics. 2003: 470-479
[4] 张海俊. 基于主体的自主计算研究[D]. 北京：中国科学院研究生院，2005
[5] 廖备水. 基于 PDC-Agent 的面向服务的自治计算研究[D]. 杭州：浙江大学计算机学院，2006
[6] 胡军. 面向自治计算的基于政策的多 agent 协同体系研究[D]. 杭州：浙江大学计算机学院，2006
[7] Wickler G J. Using expressive and flexible action representations to reason about capabilities for intelligent agent cooperation [D]. Edinburgh, U K: University of Edinburgh, 1999
[8] Sycara K, Widoff S, Klusch M, et al. LARKS: dynamic matchmaking among heterogenous software agents in cyberspace [J]. Autonomous Agents and Multi-agent Systems, 2002, 5 (2): 173-203
[9] Arisha K, Kraus S, Ozcan F, et al. IMPACT: the interactive Maryland platform for agents collaborating together [J]. IEEE Intelligent Systems, 1999, 14 (2): 64-72
[10] 蒋运承，史忠植. QoS 驱动的主体服务匹配[J]. 小型微型计算机系统，2005，26(4)：687-692
[11] 史忠植，蒋运承，张海俊，等. 基于描述逻辑的主体服务匹配[J]. 计算机学报，2004，27(5)：625-635
[12] 胡军，高济，李长云. 多主体系统中基于本体论的服务相容匹配机制[J]. 计算机辅助设计与图形学学报，2006，18(5)：694-701
[13] Hu J, Gao J, Zhou B, et al. Ontology based agent services compatible matchmaking mechanism [C] // Proceedings of the 3rd International Conference on Machine Learning and Cybernetics. Shanghai, 2004: 111-116
[14] 裘江南，仲秋雁，崔彦. 服务匹配模型中综合语义匹配方法研究[J]. 大连理工大学学报，2007，47(6)：914-919
[15] Zeng L Z, Benatallah B, et al. QoS-aware middleware for web services composition [J]. IEEE Transactions on Software Engineering, 2004, 30(5): 311-327